ISCR

Institute for Scientific Computing Research

# University Collaborative Research Program – Annual Research Reports

# University Collaborative Research Program — Annual Research Reports

The University Relations Program (URP) at the Lawrence Livermore National Laboratory (LLNL) fosters collaboration between LLNL researchers and faculty from campuses of the University of California (UC) that have the potential for unique collaborations. Major objectives of the University Collaborative Research Program (UCRP) component of the URP are to encourage original work that has the potential to significantly impact research in areas of LLNL missions and to train future Laboratory employees and faculty members with specialization in these mission areas. The ISCR portion of the UCRP program provides support for graduate students (and sometimes short-term support for postdoctoral researchers) at any UC campus. It also provides an opportunity for graduate students to interact directly with Laboratory researchers through visits and joint activities. This section summarizes the progress on ISCR's UCRP Research Grants for FY 2003.

Name, Affiliation Page

# Scalable Algebraic Domain Decomposition Preconditioners

Randolph E. Bank
Principal Investigator
UC San Diego

Shaoying Lu,
Student,
UC San Diego

Charles H. Tong
Collaborator
LLNL

Panayot S. Vassilevski,
Collaborator
LLNL

## Summary

Many leading edge scientific and engineering simulations expend large amounts of computational resources for the solution of linear systems of equations. Multilevel and domain decomposition methods have been identified as potentially scalable linear system solvers on terascale computer platforms. However, interprocessor communication overheads, degree of parallelism (in solving the coarse problems), and effects of increasing number of processors on convergence rates, all contribute to the list of obstacles to true scalability. The novel feature of our domain decomposition approach is that the subproblem residing in each processor is defined over the entire domain, although the vast majority of unknowns for each subproblem are associated with the subdomain owned by the corresponding processor. This feature ensures that a global coarse description of the problem is contained within each of the subproblems. The advantages of this approach are that interprocessor communications are minimized in the solution process while optimal order of convergence rates is preserved, and the speed of local subdomain solves can be maximized using the best existing sequential algebraic solvers.

This procedure is similar in philosophy to the parallel adaptive mesh refinement paradigm introduced by Bank and Holst, except that the present project deals with an algebraic version of the Bank-Holst paradigm in the sense that, instead of mesh refinement here, we coarsen the degrees of freedom (or matrix) outside the prescribed subdomain. Thus, the domain decomposition method applied in each processor involves the local subdomain plus a small coarse space defined on the whole domain outside the processor, and each solver utilizes only a single interprocessor communication to retrieve the global vector. This approach can be applied to general sparse matrices, although matrices arising from discretization of partial differential equations are the principal target.

Our parallel preconditioning algorithm for solving the sparse linear system $Ax = b$ is defined as follows:

*Algorithm: FocusDD solver*

$$f_p = \pi'_p * f, \text{ performed in parallel for } p = 0, 1, \cdots, P - 1$$

$$A_p = \pi'_p * A * \pi_p , \text{ performed in parallel for } p = 0, 1, \cdots, P - 1$$

$$\text{solve} \quad A_p * x_p = f_p, \text{ performed independently for } p = 0, 1, \cdots, P - 1$$

$$x = \text{collection of } \pi_p * x_p,$$

where $P$ is the number of processors used. This algorithm uses a set of rectangular matrices $\{\pi_p\}$ (prolongation operators) and a set of coarse matrices $\{A_p\}$ ($A_p$ and $\pi_p$ are stored in processor $p$), which are to be constructed and stored in sparse parallel matrix format. $f_p$ and $x_p$ are the coarsened right hand side and coarsened solution, respectively, for processor $p$.

*Continued*

Randolph E. Bank
Principal Investigator
UC San Diego

Shaoying Lu,
Student,
UC San Diego

Charles H. Tong
Collaborator
LLNL

Panayot S. Vassilevski,
Collaborator
LLNL

## *Summary continued*

For the year of 2003, we focused on fine-tuning the parallel algebraic multigrid (AMG) solvers, implemented in the software package FocusDD. In addition to our earlier two-level solvers, we have developed several new three-level solvers. The best performing among them is constructed as an additive Schwartz method, with each major iteration composed of both three-level and two-level solution steps. This FocusDD solver has been used as a preconditioner for the Generalized Minimal Residual (GMRES) method. The scalability of this solver has been studied, in comparison to that of the BoomerAMG solver, on the 16-dual-processor Beowulf cluster in the Scientific Computing Group at UC San Diego

We are preparing a report entitled *Scalable Parallel Algebraic Multilevel Methods* describing this work. Shaoying Lu will present a lecture on this work at the SIAM Conference on Parallel Processing for Scientific Computing, to take place in San Francisco on February 25–27, 2004.

We conclude with a short example. The 3D Poisson equation is solved with the FocusDD solver. A scalability study is conducted with varying number of processors (np) and a fixed number of unknowns per processor (npp). With np ranges from 2 to 128 and npp = 1000, the convergence rate and the computing time of our FocusDD solver is compared with those of the BoomerAMG solver in HYPRE. As shown in table 1, the FocusDD solver is less efficient than BoomerAMG with less than 64 processors, in both initialization and solution stages. However, the solution time of both solvers becomes comparable for 64 processors. Furthermore, the FocusDD solver becomes more efficient than the BoomerAMG with 128 processors. This trend indicates that the FocusDD solver may have better parallel efficiency than the BoomerAMG solver in the solving stage, for large clusters.

| np | FocusDD | | | | BoomerAMG | | | |
|---|---|---|---|---|---|---|---|---|
| | iter | γ | setup time | solve time | iter | γ | setup time | solve time |
| 2 | 4 | 0.026889 | 0.85 | 0.09 | 1 | 0.0 | 0.07 | 0.05 |
| 4 | 4 | 0.028566 | 1.33 | 0.10 | 1 | 0.0 | 0.12 | 0.10 |
| 8 | 7 | 0.098530 | 8.47 | 0.53 | 1 | 0.0 | 0.42 | 0.25 |
| 16 | 7 | 0.089256 | 14.90 | 0.82 | 1 | 0.0 | 1.56 | 0.60 |
| 32 | 8 | 0.164202 | 17.73 | 1.03 | 1 | 0.0 | 3.65 | 1.08 |
| 64 | 8 | 0.161620 | 89.40 | 3.43 | 1 | 0.0 | 16.45 | 4.83 |
| 128 | 8 | 0.167599 | 161.51 | 3.85 | 1 | 0.0 | 159.66 | 17.55 |

Table 1: Convergence behavior of the FocusDD solver

# Computationally Efficient Example-Based Image Segmentation

Serge Belongie
Principal Investigator
UC San Diego

Josh Wills
Student,
UC San Diego

Sameer Agarwal
Student
UC San Diego

Imola K. Fodor
Collaborator
LLNL

## *Summary*

The proliferation of digital cameras, both still and video, has produced massive amounts of visual data everywhere—from video surveillance stations to hard disks from personal low-cost digital cameras. Effective tools for image segmentation offer the potential to intelligently organize and query such collections.

Image segmentation is the problem of partitioning the pixels in an image into a relatively small number of regions that correspond to objects or parts of objects. It is one of the hardest (and oldest) open problems in computer vision and plays an important role in the process of object detection and recognition. As challenging and computationally intensive as image segmentation is, it also happens to be a problem that the human visual system solves effortlessly.

The goal of this project is to develop methods for image and video segmentation with an emphasis on motion-based processing. In particular, our two main interests are (a) image sequences containing objects undergoing large inter-frame motion (e.g., more than 10% of the image width) and (b) footage of objects in motion exhibiting temporal periodicity (e.g., a pedestrian or a running dog). Examples of practical applications likely to benefit from the proposed research include:

- tracking people, vehicles, and animals in surveillance video.
- developing vision systems for the visually impaired (e.g., a digital “seeing eye dog”).
- locating television shows based on content (e.g., “find me a nature show on cheetahs”).

Since beginning work on the project, we have developed several new algorithms that have led to promising results in a variety of problem domains involving segmentation and/or clustering. We also produced a number of publications that have been accepted or are under review.

The paper “Segmentation by Example” [1] follows directly from the proposed research activity. In this work, we operationalized the idea of inheriting local graph connections from human-labeled examples of segmented images and applied it to real images. We currently have it working on static images and are extending it to video sequences. Over the last year, several works have appeared from other research groups on learning-based approaches to image segmentation, and there appears to be a great deal of interest in this area among computer vision researchers.

Our explorations thus far in video segmentation have led to two papers — “What Went Where” [4] and “A Feature-based Approach for Determining Dense Long Range Correspondences” [5]—that extend the layer-based segmentation framework to sequences with inter-frame motion far beyond the capability of any previously existing

*Continued*

Serge Belongie
Principal Investigator
UC San Diego

Josh Wills
Student,
UC San Diego

Sameer Agarwal
Student
UC San Diego

Imola K. Fodor
Collaborator
LLNL

algorithm. Josh Wills gave the talk for "What Went Where" at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) in June 2003. The second paper is currently in review for the 2004 European Conference on Computer Vision (ECCV2004). One application area for this line of work is the deletion of unwanted objects in a series of photographs, e.g., in virtual set design for special effects.

Finally, a clustering algorithm (due to Hochbaum and Shmoys) that we encountered while pursuing the "segmentation by example" framework turned out to be ideally suited to a problem in computer graphics that was brought to our attention by Prof. Henrik Wann Jensen of UC San Diego. We pursued the application of this algorithm to the problem of accelerating image-based rendering using environment maps and the resulting paper [2] was accepted for a talk delivered by Sameer Agarwal at the 2003 meeting of the Special Interest Group in Computer Graphics (SIGGRAPH).

We are pursuing extensions of the above approaches in motion segmentation to the problem of tracking approximately temporally periodic objects in surveillance footage (e.g., pedestrians, a running dog, a bird flapping its wings). We submitted a paper called "Structure from Periodic Motion" [3] with preliminary results in this area to ECCV 2004.

## Publications

1. Sameer Agarwal and Serge Belongie, *Segmentation by example*, UC San Diego, Technical Report CS2003-0762 (2003).

2. Sameer Agarwal, Ravi Ramamoorthi, Serge Belongie, and Henrik Wann Jensen, *Structured importance sampling of environment maps*, in SIGGRAPH, San Diego, CA (2003).

3. Serge Belongie and Josh Wills, *Structure from periodic motion*, UC San Diego, Technical Report CS2003-0767 (2003).

4. Josh Wills, Sameer Agarwal, and Serge Belongie. "What went where," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognition*, volume 1, (2003), pages 37–44

5. Josh Wills and Serge Belongie, *A feature-based approach for determining dense long range correspondences*, UC San Diego, Technical Report CS2003-0768 (2003).

# Fast Direct Solvers for Large, Dense-Structured Matrices

Ming Gu
Co-Principal Investigator
UC Berkeley

S. Chandrasekaran
Co-Principal Investigator
UC Santa Barbara

Timothy Pals
Student
UC Santa Barbara

Daniel White
Collaborator
LLNL

## *Summary*

Our award is primarily concerned with the design of fast direct solvers for large, dense-structured matrices. Classical examples of large, dense-structured matrices include Toeplitz and discrete Fourier matrices. However, the kinds of matrices we are concerned with typically arise in the solution of partial differential equations and their associated integral equations. In particular they arise from the direct discretization of integral equations (including those of scattering theory) and also as the Schur complements of finite element or finite difference discretizations of partial differential equations. These are large dense matrices with complicated structures. The challenge has been to harness these structures efficiently to construct fast solvers and effective preconditioners.

Indeed a great number of algorithms have been developed to exploit such matrix structures over the last 15 years, starting with the Fast Multipole Method (FMM) of Rokhlin and his colleagues. The main innovation was to replace large subblocks of the dense matrices by low-rank approximations to significantly speed up matrix-vector multiplications. This enabled the use of iterative techniques like the conjugate-gradient method and the Generalized Minimal Residual (GMRES) method to solve the dense system of equations rapidly.

However, the speed of convergence of iterative methods is problem dependent. In fact, iterative methods can be very slow to converge or may not converge without the availability of a good preconditioner. For example, for near resonant and strong multiple scattering problems, a large number of iterations are usually required for convergence. There is now a whole area of research that looks into whether many existing integral equations can be rewritten in forms that require much fewer iterations for iterative methods. So far, some limited success has been achieved.

To get some appreciation of the complexity of the matrix structure involved, we consider the matrix whose $(i,j)$th entry is $\log|x_i-x_j|$, where the $x_i$ are points that are evenly (homogeneously) distributed between 0 and 1. Each off-diagonal block of this matrix can be approximated by a submatrix whose rank does not exceed some small number depending on the precision ($< 10$ for single precision). Other partitions of this matrix are also possible depending upon the expansions used. However, the structure becomes even more complicated in 2D and 3D. Even points distributed homogeneously on a curve in two-dimensional space can lead to a complex structure if the curve has a complex geometry.

We showed how a $ULV^T$ factorization could be used to design fast backward stable algorithms for this class of problems in one dimension. Our success is based on three novel ideas: First, we developed a new algebraic matrix structure called a sequentially semi-separable algebraic structure; basically, this is the matrix structure of the inverse

*Continued*

Ming Gu
Co-Principal Investigator
UC Berkeley

S. Chandrasekaran
Co-Principal Investigator
UC Santa Barbara

Timothy Pals
Student
UC Santa Barbara

Daniel White
Collaborator
LLNL

*Summary continued*

of semi-separable matrices and versions of it were discussed by Dewilde and van der Veen in the context of time-varying linear systems. Then we showed that linear systems of equations involving such matrices could be solved in $O(n\, r^2)$ flops, where $n$ is the matrix dimension and $r$ is the rank of off-diagonal blocks. Finally, we showed that the coefficient matrices arising from discretization of certain integral equations could be compressed into this form with an efficient compression algorithm. While this algebraic structure is specifically designed for problems in one spatial dimension, it can be applied to any matrix, even those that arise from problems with three spatial dimensions.

We showed that the peak off-diagonal ranks are quite low for matrices that arise from the global spectral discretization of the two-dimensional exterior scattering problem using the method of Kress. This method is of particular interest to our LLNL collaborators and their EIGER project. Again, this is a very high-order spectral method, and there was no fast matrix-vector multiplication algorithm even for the associated matrix. Again, as can be seen our one-dimensional algebraic approach captures the structure very efficiently and produces a fast direct solver too. This was a very pleasant surprise. Part of our proposed work is to work out a theoretical reason for the surprising effectiveness of the sequentially semi-separable algebraic structure.

We have implemented and tested the algorithm in several cases and have had good results with timing and accuracy [1,2,6]. Our numerical experiments show that our algorithm is a couple of orders of magnitude faster than standard methods, for practically important ranges of the peak off-diagonal rank.

Many of the examples we tested include those provided by our LLNL collaborators. Our numerical experiments show that the sequentially semi-separable structure is even doing a good job on problems with three spatial dimensions.

We have investigated $ULV^T$ factorization based fast algorithms for two-dimensional structures [5]. In particular we have produced an algebraic characterization of such matrices, and a general fast solver. The algebraic characterization is obtained by viewing the sequentially semi-separable structure as being defined on a unary tree (every node has at most one child).

As is evident we have emphasized the production of codes that can be used to estimate the actual run-times of all our algorithms. That is, we are not implementing our algorithms in Matlab. Obviously the downside of this approach is that as we start working with higher dimensional problems it takes a significant amount of work to produce good codes. Our LLNL collaborators, who have been providing us with example matrices, have helped us tremendously.

*Continued*

Ming Gu
Co-Principal Investigator
UC Berkeley

S. Chandrasekaran
Co-Principal Investigator
UC Santa Barbara

Timothy Pals
Student
UC Santa Barbara

Daniel White
Collaborator
LLNL

*Summary continued*

The thesis work of Timothy Pals, a graduate student being supported on this project, has already produced some outstanding results. Pals developed the first high-order technique for solving two-dimensional scattering problems that is amenable to acceleration by the fast multipole method. He also developed the first stable version of the fast multipole method for the two-dimensional scattering problem and invented some fast direct solvers that are directly tuned for two-dimensional scattering problems.

## Publications

1. S. Chandrasekaran and M. Gu, *Fast and stable algorithms for banded plus semi-separable matrices*, submitted to SIAM J. Matrix Anal. Appl. (2000).

2. S. Chandrasekaran and M. Gu, *A fast and stable solver for recursively semi-separable systems of equations*, in *Structured matrices in mathematics, computer science and engineering, II*, edited by Vadim Olshevsky, in the Contemporary Mathematics series, AMS publications (2001).

3. S. Chandrasekaran and M. Gu, "Fast and Stable Eigendecomposition of Symmetric Banded plus Semi-separable Matrices," 1999, *Linear Algebra and its Applications*, Volume 313, Issues 1-3, 1 (July 2000), pages 107–114.

4. S. Chandrasekaran and M. Gu, *A Divide-and-Conquer Algorithm for the Eigendecomposition of Symmetric Block-Diagonal Plus Semiseparable Matrices*, 1999, accepted for publication in *Numerische Mathematik*.

5. S. Chandrasekaran and M. Gu, *Superfast Nested Dissection* (in preparation 2004).

6. S. Chandrasekaran, P. Dewilde, M. Gu, T. Pals, and A.-J. van der Veen, *Fast Stable Solvers for Sequentially Semi-separable Linear Systems of Equations*. Submitted to SIMAX (2002).

# Numerical Study of Coexisting Superconductivity and Ferromagnetism

W. E. Pickett
Principal Investigator
UC Davis

Alan Kyker
Student
UC Davis

F. Gygi
Collaborator
LLNL

## *Summary*

Superconductivity and ferromagnetism are the two most evident macroscopic manifestations of quantum mechanical behavior, but these macroscopic quantum states are strongly antagonistic. On several grounds, it has been expected that they could never coexist; yet recently three examples have been identified ($UGe_2$, $ZrZn_2$, and URhGe). The scientific questions are several and basic: how does the supercurrent cope with the (frozen in) magnetic flux, which it normally abhors; how does the superconducting pairing arise; what is the character of the low-energy excitations? Almost 50 years ago Vitaly Ginzburg, who shared the 2003 Nobel Prize in Physics, concluded that electrodynamic considerations precluded coexistence of ferromagnetism and superconductivity. His considerations did not take into account the possibility of a "spontaneous flux phase," which is the arrangement that allows the supercurrent to tolerate the magnetic flux.

During the initial year of this grant, it has been established that—at the Ginzburg-Landau level (the same Nobel laureate Ginzburg) involving minimization of the free energy of the system near the superconducting critical temperature—an intrinsic magnetic flux is equivalent to an externally applied field. Minimization of the free energy is accomplished by solving second-order n on-linear differential equations for the complex superconducting order parameter $\vec{A}(\vec{r})$ and the electromagnetic vector potential $\psi(\vec{r})$, for which a steepest descent algorithm is used.

A second basic question concerns how the spin degree of freedom of the constituents of the pair is affected by the magnetization and involves the superconducting gap equation itself. The Fulde-Farrell-Larkin-Ovchinnikov (FFLO) solution arose from treatment of the gap equation in the presence of (magnetic) exchange splitting, which leads to new solutions that correspond to an inhomogeneous superconducting order parameter. This part of the project is involved with determining how the electron-dispersion relation (energy vs. momentum) impacts the formation of the FFLO phase.

Each of the simulations can he controlled by a graphical user interface (GUI), which graduate student Alan Kyker put together using C code and the OpenGL graphics library. A high quality and versatile visualization capability is *essential* for this project, because there is simply no alternative to direct visualization for understanding the character of the solution. The visualization code has two very useful features (besides the obvious): (1) it refreshes the image periodically (user selectable from 1/60 sec) while the equations are being iteratively solved, so the progress of the algorithm can be assessed in real-time; (2) the program provides several 3D graphical real-time-selectable views of various aspects of the simulation including magnetic field $\vec{B}$ field and magnitude/phase of the complex order parameter, with orientations controlled by the mouse.

*Continued*

W. E. Pickett
Principal Investigator
UC Davis

Alan Kyker
Student
UC Davis

F. Gygi
Collaborator
LLNL

## Summary continued

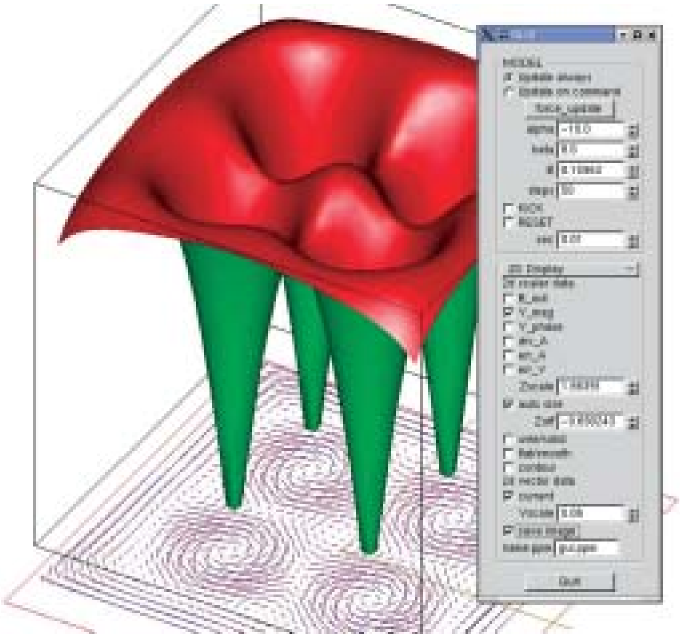

Fig. 1. Surface plot of the magnitude of the (complex) superconducting order parameter for a superconducting square wire from the code *vortex*, showing four magnetic vortices that are forced upon the system by the magnetism. Bottom: circulation of the superconducting current around each vortex. Superimposed at right is the GUI that allows easy control of parameters and re-execution (in actual applications, it does not overlay the surface plot).

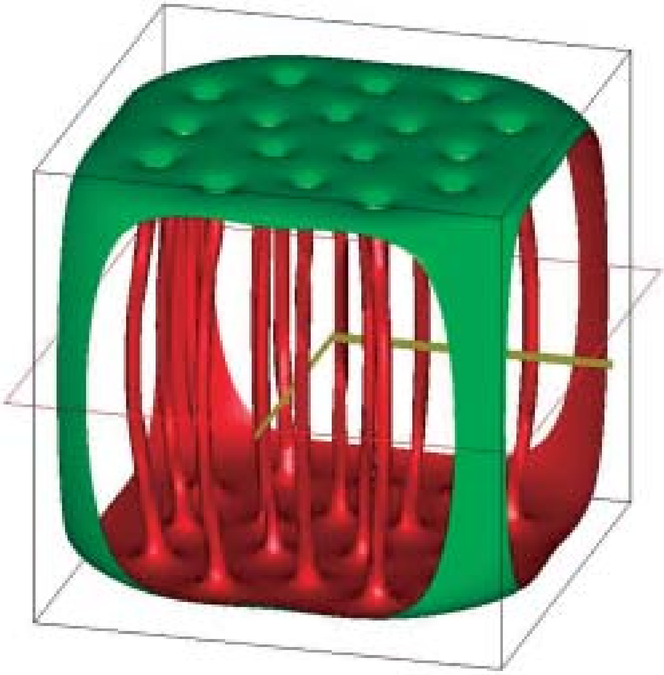

Fig. 2. Isosurface plot of the magnitude of the superconducting order parameter for a superconducting ferromagnetic cube. The spontaneous vortex lattice is evident, approaching the hexagonal arrangement predicted long ago for conventional superconductors by Alexei Abrikosov, co-recipient of the 2003 Nobel Prize in Physics.

# Simulation Of Compressible Reacting Flows

S. Sarkar
Principal Investigator
UC San Diego

A. W. Cook
Collaborator
LLNL

## *Summary*

A fundamental problem of interest regarding design of target capsules for the National Ignition Facility (NIF) is the behavior of a burn-front when it propagates through a DT mix contaminated by inert shell material introduced by Rayleigh-Taylor instabilities. In order to accurately describe burn-front propagation, we have developed a scheme that is high-order on arbitrarily non-uniform grids. Large computational grids are required to resolve the turbulence and burn chemistry. Long-time evolution is required to achieve converged statistics. Therefore, the simulations, run on the IBM Power4 machine, have stringent computational requirements. Each case is run on 128 CPUs and requires approximately 3,400 CPU hours.

The unsteady, three-dimensional form of the compressible equations for a reactive mixture of fluids is solved. A non-uniform grid is used in the *x*-axis, the direction of flame propagation, with clustering of points in the burn region. Derivatives in the *x* -direction are computed using the 6th-order compact scheme valid for arbitrary non-uniform grids. The code is parallelized using message-passage interface (MPI) and runs on the unclassified IBM SP machine at LLNL. We use a large computational box of size $680\delta_F \times 340\ \delta_F$ with $\delta_F$ denoting the flame thickness and a grid with 2304 × 1536 grid points. The integral length scale of the fluctuations is $l/\delta_F \approx 50$. Approximately 400,000 iterations are required to generate a time record that is sufficiently long for converged statistics.

In the zones contaminated by inert material, the temperature rise due to the burn-energy release is smaller than that in a pure DT mix leading to a lower reaction rate and a lower local-burn velocity. A mechanism for distorting the burn-front that is operative even in an uncontaminated mix is the so-called hydrodynamic instability, also called the Darrieus-Landau (DL) instability, that causes large-scale undulations of the front. Finally, Rayleigh-Taylor instabilities and associated turbulence in the DT mix cause wrinkling of the front.

At any given time, a burn velocity, $U_F\ (t)$, based on the overall reactant consumption rate can be defined. A time-average burn velocity, $\overline{U}_F$, can also be defined to characterize the long-time behavior of the burn-front. We have performed simulations with different levels of composition fluctuations,

$$Y' = 0, 0.02, 0.06, 0.18,$$

where $Y'$ is the root-mean-square (rms) fluctuation in fuel-mass fraction. Velocity fluctuation levels measured by the rms velocity, $u'$, have been varied,

$$u'/S_L = 0, 0.45, 1.1, 2.5,$$

corresponding to low-moderate turbulence levels. Here $S_L$ is the burn velocity for laminar, one-dimensional propagation.

*Continued*

S. Sarkar
Principal Investigator
UC San Diego

A. W. Cook
Collaborator
LLNL

## *Summary continued*

Figure 1 shows the time history of the burn velocity in a mix with *zero* imposed velocity fluctuations. For a pure reactant mix, the evolution is smooth and the time-average burn velocity, $\overline{U}_F \approx 1.3$. The natural DL instability is responsible for the 30% increase of burn-front area and associated burn velocity. Figure 1 shows that, with increasing compositional fluctuations, the burn propagation becomes progressively unsteady and the time-averaged velocity also increases.  Visualizations (not shown) suggest that the pockets of inhomogeneity seed strong DL instabilities that increase the burn-front area.  Apparently, this effect dominates the competing trend of a decrease of local-burn velocity with increasing compositional inhomogeneity so that there is a net *increase* of burn velocity with increasing compositional fluctuation.

In general, one would expect both compositional and velocity fluctuations due to RT instabilities in the reactant mix. We have varied the compositional fluctuation level for different choices of turbulence levels. Figure 2 compares the behavior of the time-averaged burn velocity, $\overline{U}_{F,}$ at a low turbulence level, $u'/S_L = 0.45$, with that at zero-velocity fluctuation. A number of conclusions can be drawn. First, for the pure reactant mix ($Z' = 0$), the burn velocity significantly increases when there are velocity fluctuations in the mix. Second, although mixture inhomogeneity, i.e., nonzero $Z'$, generally tends to increase the burn velocity, its influence is weaker when velocity fluctuations are simultaneously present in the reactant mix. Third, at high levels of mixture inhomogeneity, the burn velocity appears to be relatively insensitive to the level of velocity fluctuations.

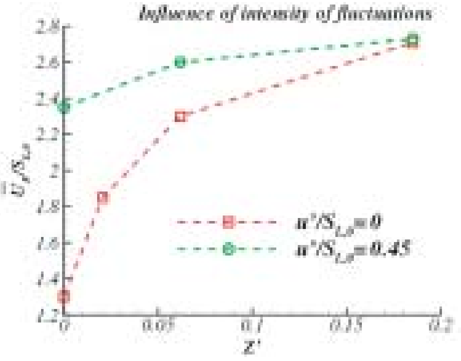

Fig. 1 The time evolution of the burn velocity.

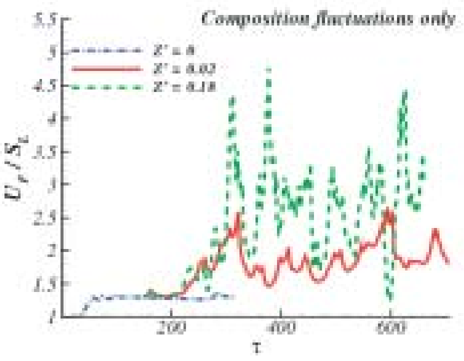

Fig. 2 The time-averaged burn velocity for different cases.

## Publications

D. G. Lopez and S. Sarkar,  "Effect of imperfect mixing on flame propagation,"*Bulletin of the 56th Annual meeting of the American Physical Society*, Division of Fluid Dynamics, New Jersey, 48 (10), 169  (2003).

# Probabilistic Clustering of Dynamic Trajectories for Scientific Data Mining

Padhraic Smyth
Principal Investigator
UC Irvine

S. Gaffney
Student
UC Irvine

Chandrika Kamath
Collaborator
LLNL

*Summary*

In this research project we have developed a set of flexible methods and algorithms for tracking and clustering time-trajectories of coherent structures (such as cyclones) in spatio-temporal grid data. These algorithms and software provide a basic set of data- analysis tools for exploration and modeling of dynamic objects, in a manner analogous to the much more widely available techniques for clustering of multivariate vector data such as k-means, Gaussian mixtures, hierarchical clustering, and so forth.

Traditional clustering algorithms typically assume that the $N$ data objects to be clustered either exist as fixed-dimensional vectors or that a matrix of $N^2$ distances or similarities is available. When applied to trajectory data (e.g., latitude-longitude positions of a storm center as a function of time) both approaches have disadvantages. The first approach requires that all trajectories be converted to fixed-dimensional vectors of the same length; this is often an unnatural representation of the data, and the underlying spatial and temporal information is no longer explicitly represented. The second approach is also problematic in that a natural distance measure may be difficult to construct; furthermore, the $O(N^2)$ computational complexity required to construct such a matrix may be intractable for large values of $N$.

To address these problems we have investigated a probabilistic alternative to trajectory clustering that uses finite mixture models. This is based on a generative model for the trajectory data in the form of a mixture of regression "shapes." Under this model, each trajectory is assumed to be generated by one of $K$ mixture components, and each component represents a mean shape (modeled as a polynomial or spline function) to which noise is added to generate an observed trajectory. The clustering problem (from a statistical viewpoint) is to recover both the parameters describing the $K$ underlying shapes (regression coefficients) and a probability distribution over shapes (or clusters) for each trajectory. We developed a set of unsupervised learning algorithms based on the Expectation-Maximization (EM) procedure that can solve this clustering problem by maximizing the likelihood of the data (multiplied by an appropriate prior). We found that this technique for clustering curves tends to systematically produce models with better out-of-sample predictive power than models that vectorize the data (Gaffney, Robertson, and Smyth, 2001).

In the methodology described above, each trajectory within a cluster shares the same regression coefficients; they only differ from each other (from the model's viewpoint) in terms of the additive noise added to each. In Gaffney and Smyth (2003) we generalized this technique such that trajectories within a cluster are allowed some random deviations (in parameter space) from the overall mean shape for that cluster. The resulting learning problem is also based on EM, and in effect it is clustering the observed trajectories in

*Continued*

Padhraic Smyth
Principal Investigator
UC Irvine

S. Gaffney
Student
UC Irvine

Chandrika Kamath
Collaborator
LLNL

*Summary continued*

parameter space, while simultaneously estimating the parameters for each curve. In experiments on simulated curves and tracked cyclone trajectories we found that this method tended to produce models with better out-of-sample predictive power than the earlier ``fixed coefficient" method (Gaffney and Smyth 2003; Gaffney, 2004).

In another extension of this work, we developed techniques to handle the real-world problem of trajectories that are mismatched in time. For example, a tracking algorithm might miss the early part of a cyclone's "life" relative to a different cyclone that is picked up by the tracking algorithm very early in its lifecycle. The resulting latitude-longitude trajectories will not be aligned, making it less likely that trajectories will be clustered appropriately. Traditional approaches involve relatively simple global alignment followed by clustering. In Chudova, Gaffney, and Smyth (2003) we developed a new technique that allows each trajectory to have an unobserved "time-offset" parameter relative to the canonical life cycle for its cluster. We showed that the EM procedure can be used to derive a statistical learning algorithm that simultaneously estimates (a) the trajectory shape for each cluster, (b) a probability distribution over clusters for each trajectory, and (c) an estimate of the most likely time-offset for each trajectory. The resulting model again shows consistently better out-of-sample performance in systematic experiments than the earlier methods.

We applied our methodology to the problem of tracking and clustering of extra-tropical cyclones (ETCs) in the North Atlantic. Understanding ETC trajectories is scientifically important for understanding both the short-term dynamics and long-term variation of atmospheric processes, such as identifying how ETC frequency and intensity distributions may be related to global climate change.

In collaboration with Dr. Andrew Robertson (IRI [International Research Institute for Climate Prediction], Columbia) and Professor Michael Ghil (UCLA) we used data from the National Center for Atmospheric Research (NCAR) Community Climate Model (CCM3) general circulation model (GCM), run with observed sea-surface temperatures specified at the lower boundary over the 1980–1995 period. For the tracking, we used the atmospheric pressure at mean sea level (MSLP) given on an approximate 2.8°× 2.8° Gaussian grid over the globe. The winds at 200 hPa (about 10-km elevation) were also available for use in interpreting the resulting cyclone clusters. The data are available every 6 hours and we analyzed data for the winter months (1 November to 30 April) from 1980 to 1995. We used a simple tracking algorithm to detect 614 cyclones of different durations, each with a minimum of 10 observations (i.e., at least 2.5 days long). From the resulting set of ETC trajectories we demonstrated that our clustering methodology was able to reveal robust and physically meaningful clusters of ETC patterns and behavior (Gaffney, Robertson, and Smyth, 2001; Gaffney et

*Continued*

Padhraic Smyth
Principal Investigator
UC Irvine

S. Gaffney
Student
UC Irvine

Chandrika Kamath
Collaborator
LLNL

*Summary continued*

al. 2004). For example, we found that cyclones tended to group into vertically, diagonally, and horizontally oriented paths in the Atlantic, with the horizontal-cluster (for example) consisting of cyclones that move horizontally (west to east) across the coastline of Europe.

Publications

S. Gaffney, A. Robertson, and P. Smyth, "Clustering of extra-tropical cyclone trajectories using mixtures of regression models," in *Proceedings of the Fourth Workshop on Mining Scientific Data Sets, Seventh* ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 15–20, August 2001.

S. Gaffney and P. Smyth, "Curve clustering with random effects mixtures," in *Proceedings of the Ninth International Workshop on AI and Statistics,* January 2003.

D. Chudova, S. Gaffney and P. Smyth, "Probabilistic models for joint clustering and time-warping of multidimensional curves" in *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers, pp. 134–141, August 2003.

S. Gaffney, A. Robertson, M. Ghil, and P. Smyth, "Probabilistic clustering of extra-tropical cyclones using regression mixture models," 2004 (in preparation).

S. Gaffney, Mixture Models for Clustering Sets of Curves, Ph.D. Thesis, University of California, Irvine, 2004 (expected).

# Memory Access Pattern Signatures and Certificates of Relevance for Benchmarks

Allan E. Snavely
Principal Investigator
UC San Diego

Michael O. McCracken
Student
UC San Diego

Jeffery Vetter
Collaborator
LLNL

Bronis de Supinski
Collaborator
LLNL

## Summary

The goal of the Department of Energy's (DOE's) Scientific Discovery through Advanced Computing (SciDAC) Program's Performance Evaluation Research Center (PERC) [see http://perc.nersc.gov/main.htm] is to develop a science for understanding performance of scientific applications on high-end computer systems and develop engineering strategies for improving performance on these systems. This project is integrating several active efforts in the high-performance computing community and is forging alliances with application scientists working on DOE's Office of Science missions to ensure that techniques and tools being developed are truly useful to end users.

The Performance Modeling and Characterization (PMaC) laboratory at the San Diego Supercomputer Center (SDSC) is a participant in PERC. It is working to bring scientific rigor to the prediction of scientific application performance on current and projected high-performance-computing (HPC) platforms with a special emphasis on the performance implication of the memory hierarchy. PMaC's goal is to predict the performance of applications more accurately than traditional benchmarking methods or traditional cycle-accurate performance simulations.

This research has investigated the performance implications of memory-access patterns and useful definitions of "signature distance" between the memory-access patterns of different basic blocks from the same or different programs. The goal is to improve the accuracy and speed of the Convolution method for performance prediction. We have enhanced the functionality of the MetaSim tool for gathering memory-access-pattern signatures and have made this tool platform-independent. We have been investigating what kinds of memory-access patterns exist "in nature" and exploring the performance implications of memory-access patterns. We have developed a nomenclature and symbolic representation of memory-access patterns, leveraging previous work by Nick Mitchell.

Michael McCracken has been examining definitions of "signature distance" between basic blocks—the idea being that (possibly) basic blocks that look roughly the same in terms of memory-access patterns may perform roughly the same on a given machine. In work defining a meaningful metric for "signature distance" he has established orthogonal properties of loops including memory footprint, memory-access pattern, type and intensity of floating-point, and Instruction-level parallelism (ILP) operations that could provide "certificates of relevance" for benchmarks. In other words, he has shown early evidence that it is possible to reasonably estimate the likely performance of an application that has been profiled by MetaSim if all of the application's basic blocks are mapped to a usefully similar set of benchmark basic blocks whose performance has already been established. McCracken has published a paper [1] showing how this approach may be applied for dynamic algorithm selection guided by performance model predictions. Currently he is applying the technique to a commercial code.

Continued

Allan E. Snavely
Principal Investigator
UC San Diego

Michael O. McCracken
Student
UC San Diego

Jeffery Vetter
Collaborator
LLNL

Bronis de Supinski
Collaborator
LLNL

*Summary continued*

McCracken studied about 108 scientific loops taken from a paper in *Parallel Computing* titled *A comparative study of automatic vectorizing compilers* by D. Levine, David Callahan, and Jack Dongarra. Using the statistical analysis tool SAS he was able to derive parameterized functions to predict the performance of these loops as a function of their memory footprint, memory-access pattern, floating-point mix, and intensity and ILP on IBM Power3 and Power4 processors. A poster is in preparation for the UC San Diego Research Review.  Another paper is in preparation that will apply the modeling method to explain and predict the performance the commercial Computational Fluid Dynamics (CFD) application Cobalt (see http://www.cobaltcfd.com/).

Publications:

1. Michael O. McCracken, Allan Snavely, and Allen Malony, *Performance Modeling for Dynamic Algorithm Selection*, ICCS, Workshop on Performance Modeling and Analysis (PMA03), Melbourne (2003).

# Statistical Inference from Microarray Data with Applications in Breast Cancer Research

Mark van der Laan
Principal Investigator
UC Berkeley

Annette Molinaro
Student
UC Berkeley

Dan Moore
Collaborator
LLNL

## Summary

The advent of microarray technologies has made feasible the measurement of gene copy number and gene expression in very large numbers from a wide variety of biologic samples. It is believed that gene copy number and gene expression are important indicators of cancer progression and that an understanding of how genes are turned "on" or "off" will ultimately lead to better strategies for treatment and to a reduction in cancer incidence.

The overall goal of this collaboration is to link gene microarrays to disease progression. This goal includes the following components: to determine reliable methods for classifying tumors based on their genetic profile, to link the classified tumors to covariates and clinical outcomes, to link the gene microarrays to covariates, and to link these gene microarrays and covariates to clinical outcomes.

This year we initially focused on a data set consisting of 152 breast cancer cases, for which we had comparative genomic hybridization (CGH), pathological, and clinical covariates. Our goal was to find chromosomal loci where specific alterations would be predictive of survival time. In this project we proposed a unified strategy for estimator construction, selection, and performance assessment in the presence of censoring. This approach is entirely driven by the choice of a loss function for the full (uncensored) data structure and can be stated in terms of the following three main steps: (1) Define the parameter of interest as the minimizer of the expected loss, or risk, for a full data-loss function chosen to represent the desired measure of performance. Map the full data-loss function into an observed (censored) data-loss function having the same expected value and leading to an efficient estimator of this risk; (2) Construct candidate estimators based on the loss function for the observed data; and (3) Apply cross-validation to estimate risk based on the observed data-loss function and to select an optimal estimator among the candidates. A number of common estimation procedures follow this approach in the full-data situation, but depart from it when faced with the obstacle of evaluating the loss function for censored observations.

Tree-based methods, where the candidate estimators in Step 2 are generated by recursive binary partitioning of a suitably defined covariate space, provide a striking example of the chasm between estimation procedures for full data and censored data (e.g., regression trees as in Classification and Regression Trees [CART] for uncensored data and adaptations to censored data). Common approaches for regression trees bypass the risk-estimation problem for censored outcomes by altering the node-splitting and tree-pruning criteria in manners that are specific to right-censored data. In this project we implemented a generalization of regression trees to censored data. We used the CART (Breiman, et al. 1984) algorithm altered by the general-estimating equation methodology of Mark van der Laan in collaboration with James Robins (van der Laan & Robins 2003).

*Continued*

Mark van der Laan
Principal Investigator
UC Berkeley

Annette Molinaro
Student
UC Berkeley

Dan Moore
Collaborator
LLNL

*Summary continued*

The formalization of this approach encompasses univariate and multivariate prediction and density estimation for censored and non-censored data. This project has resulted in a technical report (www.bepress.com/ucbbiostat/paper135/) and a manuscript in press for a special issue, ”Genomics,” of the Journal of Multivariate Analysis.

Upon completion of this project, we began investigating other piecewise constant regression methods based on biological reasoning for generating candidate estimators in Step 2. CART evaluates and chooses the best binary splits of the covariates, building a list of *“and”* statements that predict survival times (e.g. loss at locus 1 *“and”* gain at locus 2 predicts survival of x months). However, due to reasoning that multiple chromosomal aberrations could lead to a similar or identical effect, CGH and expression data may require alternatives to the *“and”* statement ordering (e.g. loss at locus 1 *“or”* loss at locus 3 *“and”* gain at locus 2 predicts survival of y months). Our current project involves building a new algorithm which includes *“and”* and *“or”* statements. This algorithm is more aggressive than CART by allowing a variety of covariate splits and subsequently unions of those splits. Annette Molinaro is in the final stages of implementing this algorithm for univariate prediction in the statistical software R for distribution as an R package.

Preliminary results shown in Table 1 based on simulated data show that our algorithm describes the covariate space more aggressively (i.e., with a smaller average risk over an independent test sample [column 3]) than CART and with fewer parameters (Average Size [column 4]) than CART. This can be seen over several sample sizes (i.e., n 2 [250, 500, 1000]). All simulation results and a complete description will be available in the technical report ”A Deletion/ Substitution/ Addition algorithm for partitioning the covariate space in prediction” by Annette Molinaro and Mark van der Laan.

| n | Method | Average Risk | Average Size |
|---|---|---|---|
| 250 | ours | 0.288 | 8.5 |
|  | CART | 0.397 | 13.2 |
| 500 | ours | 0.1796 | 13.0 |
|  | CART | 0.234 | 20.4 |
| 1000 | ours | 0.137 | 15 |
|  | CART | 0.161 | 26.5 |

Table 1. Preliminary results based on simulated data.